YBS Ansiklopedi

www.YBSAnsiklopedi.com

Cilt 5, Sayı 2, Temmuz 2018

CRISP-DM: Endüstriler Arası Standart İşleme – Veri Madenciliği için (Cross Industry Standard Processing – Data Mining)

Şadi Evren ŞEKER[1]

1. *Bilkav Eğitim Danışmanlık*

Özet

Veri bilimi, hızla gelişen ve her gün yeni sektörlere ve yeni problemlere açılan günümüzdeki en önemli ve hızlı çalışma alanlarından birisidir. Veri bilimi alanında yıllar süren çalışmalar, çok sayıda farklı alanda yapılan projeler ve ihtiyaçlar, bu alanda belirli standartlaşmalara ihtiyaç doğurmuştur. Bu alanda yıllar içerisinde farklı yaklaşımlar geliştirilse de ulaşılan son noktada, bir veri bilimi projesine nereden başlanacağı, hangi adımların izlenmesi gerektiği, projenin aşamalarının çıktıları ve proje süresince ölçülebilir adımları CRISP-DM olarak kısaltılan yöntemle yönetilebilmektedir. Veri bilimi projelerinin farklı alanlara uygulanabilir olmasının yanında, alan bağımsız olarak yönetilebilir olmasını sağlayan bu yöntemin çıkışı, detayları, kullanım şekli ve günümüzdeki ulaştığı noktayı bu makalede ele aldık.

Anahtar Kelimeler: *veri madenciliği, veri bilimi, büyük veri analitiği, proje yönetimi*

Abstract

Data science is one of the most important and fast working areas now a days, it is also rapidly developing and opened up to new sectors and new problems every day. Many years of work in the field of data science, projects and needs in a number of different areas have created a need for standardization in this area. Although different approaches have been developed in this field over the years, it is possible to manage a data science project at the last point, which steps should be followed, outputs of the project stages and metrics during the project are abbreviated as CRISP-DM. In addition to the fact that data science projects are applicable to different fields, we have discussed the output, details, usage and current point of this method which makes the field independently manageable.

Keywords: *data mining, data science, big analytics, Project management*

1. Giriş

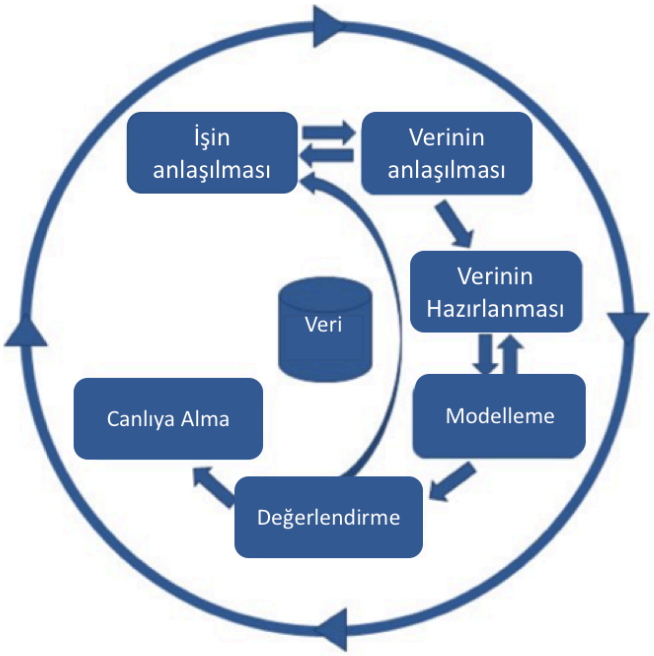
Veri bilimi projelerinin kısa tanımı: "Verinin olduğu her yerde değer üreten projeler" olarak yapılabilir. Günümüzde her yerde veri olduğunu artık kabul etmemiz gerekiyor. Kolumuzdaki saatten, cebimizdeki cep telefonuna, sürekli kayıt yapan güvenlik kameralarında, benzin pompalarına, uzaya fırlatılmış uydulardan evdeki su sayacımıza kadar hemen her şey biz farkında olalım veya olmayalım veri üretiyor, topluyor, iletiyor veya işliyor. Toplanan devasa boyuttaki veriler bizim için saklama, işleme veya iletme gibi problemler olarak karşımıza çıkıyor. Bunları doğru işleyen veya akıllı sistemler geliştirerek akıllı sistemlerin işlemesini sağlayanlar ise verinin daha

değerli hale gelmesini sağlıyor. Şu anda toplanan veri çoktan insanlığın işleme kapasitesini geçmiş durumda. Bir yılda toplanan veriyi dünyada yaşayan bütün insanların sadece okuması bile artık imkansız boyutta, bu verileri anlamak, anlamlandırmak, bu veriler üzerinden karar vermek gibi süreçler ise elbette çok daha zeki ve yüksek işlem güçleri gerektiriyor.

Bütün bu devasa problemler bizi zeki sistemler geliştirmeye ve bu sistemlerin bizim için çalışmasını zorunlu kılıyor. Bugün firmalar üçüncü veri dönüşümlerini yaşıyorlar. Buna göre önce veriyi tutan ve temel veri tabanı yönetim sistemlerine geçen firmalar, ikinci aşamada akıllı sistemlerle buluşup tanımlayıcı analitiğe giriş yaptılar, skor kartları, göstergeler (dashboards), veya iş zekası gibi kavramlar hayatımıza girdi, şimdi ise büyük veri ve yapay zeka gibi kavramlarla karşılaşan kurumlar artık üçüncü veri dönüşümünü gerçekleştiriyorlar. Bu dönüşüm elbette belirli bir sistemle yapılması gereken ve gerek proje maliyetlerinin hesaplanması gerekse dönüşümün doğru yönetilmesi için belirli metodolojilerin konulması gereken bazı problemler içeriyor. Bu yazı kapsamında günümüze kadar kullanılan farklı veri bilimi proje yönetim sistemleri olan SEMMA ve KDD yöntemleri açıklandıktan sonra CRISP-DM yönteminin farkı ve avantajları/dezavantajları tartışılacaktır. CRISP-DM'in bize kazandırdığı bakış açısı ve projelere sağladığı felsefe anlatıldıktan sonra bu yönteme karşı kritik ve bu yöntemin günümüzdeki kullanımlarından örnekler de sunulacaktır.

2. CRISP-DM Nedir

En temel tanımı itibariyle 6 adımlı bir süreçten ibarettir. Bu süreçler arasındaki geçiş aşağıdaki şekilde görselleştirilmiştir.



Şekil 1: CRISP-DM Adımları ve Akışı

Görüleceği üzere bazı adımlar arasında döngüler mümkündür ve bunların proje yönetimi açısından önemi yüksektir ancak bu döngülere ve geçişlere geçmeden önce, 6 temel adımı tanıyarak başlayalım:

1. İş Süreçlerinin anlaşılması: bu aşamada problem tanımı yapılır. Yani bir problemin neden çıktığı, problemin çözümündeki beklentiler, iş süreçlerinde problemin dokunduğu ilgili veri kaynakları ve veri akışları tespit edilir. Problemin çözümü sonunda hangi çıktıların beklendiği tanımlanır. Örneğin tarım alanında bir problem, önümüzdeki sene, Türkiye'ni buğday rekoltesinin tahmin edilmesi veya tekstil alanında çalışan bir firma için problem, hangi müşterilerin şikayeti olduğunun önceden tespit edilmesi veya Telekom alanındaki bir problem hangi müşterinin rakip Telekom firmasına geçeceğinin önceden tahmin edilmesi, veya pazarlama alanındaki bir problem, hangi müşteriye hangi reklamın gösterileceği olabilir.

2. Verinin anlaşılması: ikinci adımda, probleme uygun olarak veri toplanır veya mevcut verinin üzerinden geçilir. Genelde bu adıma başlanmadan önce problem tanımının doğru yapılması oldukça önemlidir çünkü sık yapılan hatalardan birisi, eldeki verinin tamamının gereksiz yere işlenmesidir. Yine bu aşamada veri ile ilgili problemler de ortaya çıkarılır. Örneğin verinin gürültülü veya kirli olması, eksik veri içerilmesi gibi problemlerin tespit edilmesi, verinin yapısal / yapısal olmayan veya yapısal verinin tipinin tespit edilmesi bu aşamada yapılır. Gerekli görülürse ilave veri toplanır veya eldeki veriler üzerinden nasıl işlemler yapılarak verinin zenginleştirilebileceğine bu aşamada karar verilir.

3. Veri Ön İşleme aşaması: Bu aşamada veri üzerinde yapılacak işlemlere ve bu işlemlerin hangi yöntemlerle yapılacağına karar verilir. Örneğin, bir önceki adımda, veri üzerinde eksik veri tespiti yapıldı, bu tespite göre bir kısmı eksik olan verinin sisteme hiç dahil edilmemesi veya verinin eksik kısımlarının tamamlanması, bu tamamlama sırasında nasıl bir yöntem izleneceği (bkz. Töhmet (imputation) [1]) gibi kararlar bu aşamada verilir, uygulanır ve çıktılar değerlendirilerek daha başarılı hale getirilmeye çalışılır. Örneğin doğum tarihlerinin yaşa verilmesi basit bir veri dönüşümü (transformation) veya adres alanında kişinin il ilçe gibi bilgilerinin çıkarılması veri zenginelştirme (enrichment) işlemleridir. Bu aşamada yapılan çalışmalar bir önceki aşama ile birlikte ele alındığında, literatürde öznitelik mühendisliği (feature engineering) adı verilen bir başlık altında incelenebilir.

4. Model aşaması: Aslında veri analitiği projelerinin, veri madenciliğine doğrudan dokunduğu aşama bu aşamadır. Bu aşamada tanımlanan problem ve veri kaynakları üzerinde bir makine öğrenmesi veya istatistiksel model geliştirilir. Geliştirilen model, istenen problem çözümüne yönelik olarak iyileştirilir (optimization). Bu aşamada genelde modele uygun olarak verinin düzenlenmesi gerekebilir. Örneğin bazı modellerin çalışması için dengeli veri veya bazı modeller için aykırı verinin ayıklanması ve hatta başka bazı modeller için de veri dönüşümleri gerekebilir.

5. Değerlendirme aşaması: Bu aşamada, şimdiye kadar olan adımların genel bir değerlendirmesi yapılır ve aslında ilk adımda (problem tanımı aşaması) konulan başarı kriterlerini ne ölçüde sağladığı test edilir. Örneğin müşterilere ürün tavsiyesi yapan bir sistemin, doğru müşteriye doğru ürün tavsiyesi sonucunda satışları arttırması beklenir ve geliştirilen yeni sistemin, satışlarda ne kadar artış sağladığı ölçülür. Bunun için deney grupları veya farklı test tekinleri kullanılabilir. Örneğin müşterilerine doğru reklamı göstermek için veri bilimi projesi geliştiren bir e-ticaret sitesinde, veri bilimi kullanılan ve kullanılmayan iki grup oluşturarak üzerlerinde A/B testi uygulanabilir. Bazı durumlarda, saha üzerinde test imkanı olmadığı için geliştirilen yöntemin başarısı bazı metrikler üzerinden ölçülebilir. Problem tanımlarına göre örneğin kesinlik (accuracy) saflık (purity) gibi değerlere bakılabilir. Verilen karara göre ürünleşmeye geçilebilir veya bütün aşamaların tekrar gözden geçirilmesi için ilk aşamaya geri dönülebilir.

6. Ürün aşaması: Bu aşamaya kadar, elde edilen çıktılar kurumun çalışma ortamına uygun halde geliştirme sürecine başlanır. Örneğin kullanılıyorsa, büyük veri platformuna uygun halde veya kurumda kullanılan programlama dilleri ve veri akışına uygun hale getirilerek sistemin çalışan bir uygulaması geliştirilir. Örneğin bir e-ticaret sitesindeki reklamların veri madenciliği süreçleri ile yönetileceği bir sitenin kodlanmasına başlanır, reklam gösterimlerine karar verilen adımlarda ilgili değişiklikler yazılarak yeni geliştirilen karar sistemi devreye sokulur ve karar sisteminin verdiği sinyallere göre reklamlar gösterilmeye başlanır.

Şekil 1'de gösterilen adımlar arasında bazı durumlarda döngü oluşur. Örneğin veri analizi aşaması ile işin anlaşılması adımları arasında döngü bulunur. Bunun sebebi işe göre verinin toplanması ve veri toplama adımında yaşanan zorluklara veya fırsatlara göre iş analizinin yeniden gözden geçirilmesidir. Benzer bir döngü de model ile veri ön işleme adımları arasında vardır. Bunun sebebi de model seçimine göre bazı veri ön işleme adımlarının gerekli olmasıdır. Örneğin bazı modeller eksik veri ile çalışamamaktadır. Bir diğer sebep de veri ön işleme adımında elde edilebilecek bazı çıktıların modelin iyileştirilmesinde kullanılabilecek olmasıdır. Son olarak, yöntemin en önemli döngüsü değerlendirme adımındaki sonuçlara göre sistemin en başa dönerek bütün adımların üzerinden geçebilmesi anlamına gelen ana döngüdür.

3. CRISP-DM ve Uygulama Tecrübeleri

Sektörde uzun süreler CRISP-DM kullandıktan sonra, yöntemin getirdiği bir bakış açısı ve felsefe, uzmanlar tarafından kabulleniliyor. Buna göre bir veri bilimcisi çok istisna durumlar olmadıkça, bir projeye başlarken önce problem ve iş süreçlerinin analizini yapar. Bu aşama tamamlandıktan sonra verinin analizine başlanılmalıdır. Örneğin "bu veriden nasıl bir proje yaparım?" gibi sorular aslında CRISP-DM'in önünü almaya çalıştığı ve kabul görmeyen sorulardır, çünkü verinin toplanması, problem tanımından sonra gelir. Benzer şekilde sık rastlanan bir durumda veri hikaye anlatıcıları (data story tellers) ve veri kahinleridir (data fortune tellers). Hikaye anlatıcı veya kahin durumuna düşmemek için CRISP-DM tarafından tanımlanan iki adımın doğru şekilde geçilmesi çok önemlidir. Buna göre şayet problem tanımı yapılmadan veri üzerinden projeye geçiliyorsa aslında bir problem çözülmeye çalışılmıyor ama verinin hikaye olarak anlatılması (story telling) yapılıyor demektir. Benzer şekilde şayet elde veri olmadan sadece problemler üzerinden çözüm aranıyorsa ve veri aşaması zayıf geçiliyorsa, bu durumda da kehanet okunuyor demektir.

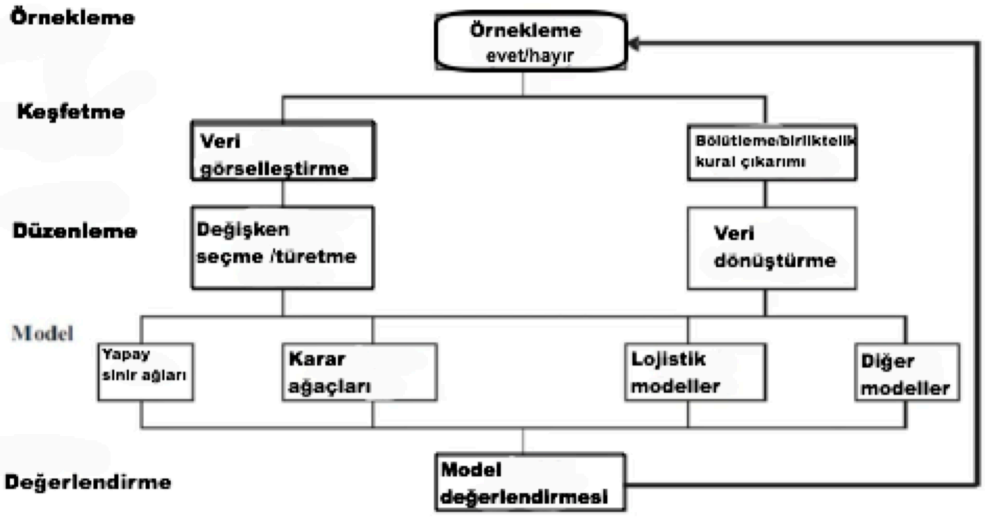
CRISP-DM ayrıca projelerin organizasyonu ve ölçülmesi için de bazı kriterleri görmemizi sağlar. Örneğin proje organizasyonunda CRISP-DM'in her adımı için aşağıdaki ünvanlarda kişilere ihtiyaç duyulabilir:

1. Problem ve iş süreçlerinin tasarımı aşamasında iş anlistleri (business analyst)
2. Verinin anlaşılması, toplanması aşamalarında veri analistleri (data anlysts)
3. Veri ön işleme aşamasında : Öznitelik Mühendisi (Feature Engineer), Veri Mühendisi (Data Engineer), Veri Analisti (Data Analyst) Veri Bilimci (Data Scientist) veya ETL Uzmanı [2]
4. Model oluşturma aşamasında: Veri Bilimci (Data Scientist), Veri Madencisi (Data Miner)
5. Değerlendirme aşamasında
6. Gerçekleme aşamasında: yazılım geliştirici (software developer), sistem operatörü (SysOp), geliştirme operatörü (DevOp)

Yukarıdaki listedeki ünvanlar, zamana göre değişebilmektedir. Ayrıca henüz yeni bir dünya olan veri biliminde, ünvanlar da tam olarak netleşmemiştir ve farklı şirket kültürlerinde, farklı ülkelerde veya bakış açılarında (istatistik kökenli birisi ile bilgisayar mühendisliği veya işletme / endüstri mühendisliği temelli kişilerin) isimlendirmeleri farklı olabilmektedir.

4. CRISP-DM Çıkışı ve Diğer yöntemler:

CRISP-DM, ilk çıkan veri bilimi projeleri yönetim metodu değildir. Daha öncesinde SEMMA ve KDD ismi verilen iki yöntemin varlığından bahsedilebilir. Kronolojik olarak ilk çıkan yöntem SEMMA yöntemidir ve adımları CRISP-DM'e oldukça benzer. Aslında SEMMA yönteminin, bir anlamda CRISP-DM için geçiş süreci olduğu ve CRISP-DM'e evirildiği söylenebilir.



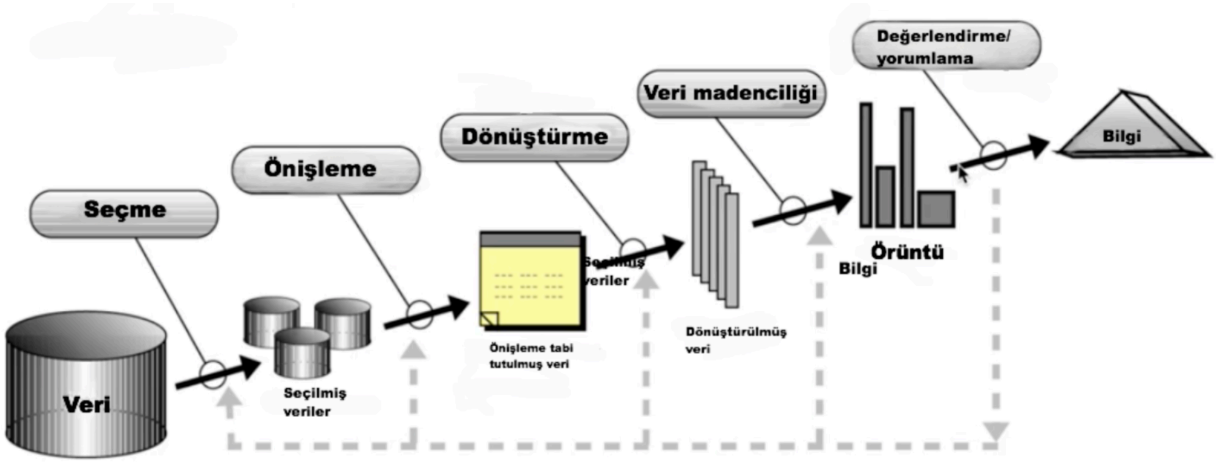
Şekil 2: SEMMA Adımları ve Akışı

SEMMA yöntemi örnekleme(sample) ile başlamaktadır. Adımları, sırası ile : Sample(Örnekleme) , Explore, (Keşfetme) Modify (Düzenleme), Model (Modelleme) ve Evaluate (Değerlendirme) olarak geçer. Adımlar ele alındığında, CRISP-DM adımlarına benzer şekilde ilerlemektedir. SEMMA'nın çıkışı, veri işlemenin kısıtlı imkanlarla yapıldığı dönemlere dayanır. Günümüzdeki büyük veri işleme kapasiteleri yerine oldukça kısıtlı işleme kapasitesi olan zamanlarda, veri üzerinden öncelikle işlenebilir miktarda örnekleme yapılır ve bu örnekleme üzerinde veri analizine başlanırdı. SEMMA, CRISP-DM'den farklı olarak, problem tanımını içermez, yani endüstrideki çoğu uygulamada değişen problemler ve bu problemler için verinin farklı kullanımları SEMMA'da bulunmaz. Verinin farklı amaçlara yönelik olarak bir seferde işlenmesi amacıyla geliştirilmiştir.

SEMMA'nın ilk adımı örneklemedir ve örnekleme aşaması yapılmadan başlanırsa verinin keşif sürecinde bölüteleme / kural çıkarımı gibi yöntemler kullanılarak veri üzerinden anlamlandırma işlemleri yapılır. Şayet keşif süreci bir örneklem üzerinden yapılıyorsa, bu durumda verinin görselleştirilebilir boyutta olduğu kabul edilerek veri üzerinde uzman bir gözün analiz ve keşif süreci başlar. Keşif aşaması geçtikten sonra veri üzerinde gerekli düzenlemeler yapılır. Bu düzenleme adımı, CRISP-DM'de bulunan veri önişleme adımına benzetilebilir.

SEMMA'nın en önemli çıktı oluşturan adımı ise model aşamasıdır ve bu adımda istenen problem tanımına uygun olarak yapay sinir ağları, karar ağaçları, lojik modeller veya veriye uygun olarak makine öğrenmesi ve istatistiksel modeller çıkarılır. CRISP-DM'e benzer şekilde modellerin değerlendirildiği son adımla SEMMA adımları tamamlanmış olur.

Benzer olarak günümüzde hala kullanılan KDD adımları da CRISP-DM gelişiminde ve algısında önemli rol oynamıştır.



Şekil 3: KDD Adımları ve Akışı

KDD literatürde "Knowledge and Data Discovery" veya "Knowledge Discovery in Data" veya "Knowledge Discvory in Databases" gibi kelimelerin kısaltılması olarak kullanılabilir. KDD proje yönetim metodunun anlaşılabilmesi için Knowledge (bilgi) ve Information (Enformasyon) kavramları arasındaki farkın anlaşılması gerekir. Özetle veri anlam ifade etmeyen sayılar ve karakterler olarak görülebilir. Bu anlamsız sayı ve karakterlerin basit bir soruya cevap vermesi durumunda enformasyona dönüştüğünü ve bu basit soruların zeki sorulara dönüştüğü anda da bilgiden bahsedilebilir. Örneğin 1000 tek başına bir sayı iken, bu sayının bir kişinin maaşı olması durumunda enformasyondan bahsedilebilir. Benzer şekilde, şirketteki maaşların ortalaması, maaşların en düşüğü veya en yükseği enformasyon seviyesine ait birer sorudur. Bu tip sorular genelde veri tabanı seviyesindeki sorgular ile çözülebilir. Bilgi seviyesine gelindiğinde ise daha zeki sorular sorulabilir. Örneğin yeni işe başlayacak birisinin mezuniyeti, iş tecrübesi, çalışacağı pozisyon gibi bilgilere bakarak maaşını tavsiye eden bir sistem bilgi (knowledge) seviyesi olduğu söylenebilir.

KDD adımları sırasıyla aşağıdaki şekilde açıklanabilir:

1. Seçme (Selection): Probleme uygun verilerin seçilmesidir.
2. Ön İşleme (Preprocessing): CRISP-DM süreçlerinde yer alan veri ön işleme aşamasıdır ancak KDD adımlarında önişleme ve dönüştürme süreçleri ayrılmıştır. CRISP-DM bu iki dönüşümü tek bir aşama olarak ele alır. Bununla birlikte KDD için ön işleme süreçleri eksik verilerin tamamlanması, kirli ve gürültü verilerin çözülmesi gibi adımları ön işlemede ele alır.
3. Dönüştürme (Transformation): Verinin dönüştürülmesi ayrı bir aşamada ele alınır. Verinin zenginleştirilmesi veya farklı tiplere ve içeriğe dönüştürülmesi bu aşamadadır. Örneğin doğum tarihlerinin yaşa çevrilmesi veya doğum tarihlerinden kişilerin burçlarını çıkarıp müşteri davranışları üzerinde burçların etkisi olduğunun araştırılması dönüştürme aşamasında ele alınır.

4. Veri Madenciliği (Data Mining): CRISP-DM aşamalarından model oluşturma aşamasına benzetilebilir. Bu aşamada istatistiksel veya makine öğrenmesi modellerinin geliştirildiği aşamadır.
5. Değerlendirme (Evaluation): Yine CRISP-DM aşamalarından değerlendirme aşamasına benzetilebilir, verinin bu zamana kadar olan yolculuğu sonucunda çıkarılan örüntülerin (pattern) yorumlandırğı ve artık bilgiye dönüştüğü son aşamada, elde edilen çıktıların değerlendirildiği aşamadır.

5. Sonuç

Günümüzde, veri bilimi çalışmalarında sıkça kullanılan CRISP-DM yöntemine bir giriş yapılmış ve bu yöntemin aşamaları ve kullanımları detaylandırılmıştır. Yine literatürde çok sık kullanılan KDD vey SEMMA yöntemleri ile CRISP-DM'in ortak ve farklı yanları ortaya konularak karşılaştırılmıştır. Çok sayıda veri bilimi ve veri madenciliği aracı, günümüzde CRISP-DM yöntemini içselleştirmiştir [3]. Örneğin SPSS Modeler gibi araçlarda doğrudan CRISP-DM adımları açılmakta ve araç, veri bilimciyi bu adımlara yönlendirmektedir.

Kaynakça

[1] Şeker, Ş. E. ; Eşmekaya, E. (2017), Eksik Verilerin Tamamlanması (Imputation), *YBS Ansiklopedi*, v. 4, is. 3, pp. 10 – 17

[2] Eşmekaya, E. ; Şeker, Ş. E. (2017), ETL Süreçleri (ETL Process), *YBS Ansiklopedi*, v. 4, is. 3, pp. 18- 34

[3] Yılmaz, M. ; Şeker, Ş. E. (2016), Veri Madenciliği Araçları (Data Mining Tools), *YBS Ansiklopedi* , v. 3, is. 4, pp. 10 – 20